International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 02, 2021

A Framework For Enhancing The Accuracy Of K-Means Clustering Algorithm With Linear Data Structures By Removing The Outliers

James Manoharan. J

Dept. of Computer Applications,Bishop Heber College (Autonomous), Tiruchirappalli, India
(Affiliated to Bharathidasan University, Tiruchirappalli)

Email:james.ca@bhc.edu.in

Abstract: Clustering is a common technique for statistical data analysis, which can be used in various fields, like data mining, machine learning,pattern recognition, bioinformatics and image analysis.It is the method of grouping associateddata objects fromdissimilarsets, and it partitionsdatasetsas subsets.So that the data object of each subset rendering to the defined distance degree. K-means is a very well-known clustering algorithm for its nature of simplicity and the power of computational efficiency. Similarity of data objects in K-means algorithm is identified using the measure of distance which leads to implement robust algorithms in both the functionalities of classification and clustering.The measures of distance play a vital role in the overwhelming performance of K-means algorithm. The crucial functionality of distance metrics is to measure the distance between data objects in a dataset.The K-means algorithm calculates the distance between the centroids and data objects. The clusters are formed by grouping the data objects to centroids with minimum distance based on the resultant values [Nasooti et al. 2015]. Therefore, the calculation of distance plays a major role in the process of clustering. Choosing a proper technique for distance calculation is totally dependent on the type of the data.

Keywords: K-means Algorithm, Linear data structure, Data object, Cluster analysis, Outlier detection

1. INTRODUCTION

Clustering is one of the classification methods to classify the given data objects and find out the hidden information which exists in given datasets[4].Clustering algorithm partitioning the given data objects into clusters, such that the data objects in one cluster are similar to each other[2]. Clustering techniques are widely applied in many application areas such as information retrieval, bio-informatics, medicine, neural networks and pattern recognition and so on. K-means clustering algorithm is one of the most popular clustering algorithms. It uses Euclidean distances to find out the distance between each data objects. K-means is the most generally used clustering algorithm in exercise. K-means clustering is an unsupervised, numerical, non-deterministic iterative technique. Its simplicity and usefulness are notable in all the available approaches.In this research work, we create a new framework for k-means algorithm with linear data structures which finds the problem of outliersi.e (irrelevant data points) and increase the efficiency of traditional k-means algorithm with Linear data structure. The framework is composed of 3 stages; choosing initial k-centroids phase, calculate the distance phase and recalculating cluster center phase. The initial process of choosing initial k-centroids phase the initial cluster centers have obtained using divide-and-conquer method. The distance phase calculation stage discovers the distance between each data objects and cluster centers in each iteration can be intended using linear data structure List.

2. RELATED WORK

Clustering is the task of assigning a set of data objects into groups called clusters in which data objects in the identical cluster are more similar to each other than to those in other clusters. In general clustering is used to discover the similar, dissimilar and outlier data items from the databases. The main idea behind the clustering is the distance between the data items [Purohit et al. 2015].

Jadwal et al. (2012)have proposed An Improved and Customized I-K Means for avoiding Similar Distance Problem. The authors have generated an approach for solving similar distance problem using improved K-means clustering. Quality factor can be described in terms of reduces the intra class similarity and maximizing the inter class similarity.

Kaur et al. (2012)have generated an Efficient K-Means Clustering Algorithm Using ranking method in Data Mining. This work has made an attempt at studying the feasibility of K-means algorithm in data mining using the Ranking Method.

Authors in [4] proposed a new method to pit the initial centers of k-means. The design is based on the perception of scattering the particular k initial cluster centroids not nearer to each of them, the initial cluster midpoint is chosen consistently by random from that data points are being grouped, after that each successive centroid is selected from the remaining data points and the probability is proportional to its distance formed to the position closer cluster'smidpoint.

Suryawanshi et al. (2015)have proposed a review paper of various enhancements for clustering algorithms in Big Data Mining. This work reviews different improvements and techniques of K-means clustering algorithm. These methods included refined initial cluster center's method; a parallel K-means algorithm and a parallel K-means clustering algorithm based on map reduce technique, find out the initial centroids of the clusters and assign each data point to the appropriate matching clusters.

3. METHODOLOGY

3.1 PERFORMANCE ENHANCEMENT OF CLUSTERING BY FINDING OUTLIERS USING K-MEANS CLUSTERING WITH LINEAR DATA STRUCTURES

In clustering, the distance between two points can be calculated in different ways. The challenging task is to choose an appropriate technique from the available ones. In fact, the selection of distance techniques also considered to be important with the property of data and the dimension. This chapter presents an enhanced K-means clustering algorithm using linear data structure list. The enhanced method that improves the efficiency of clustering by calculating the distance between the data objects in an efficient way. And it presents nearly theoretical and experimental analysis of the proposed method. The proposed K-means algorithm produces the same clustering result as obtained by the traditional K-means method but in a reduced time.

Algorithm 1: THE SIMPLE K-MEANS CLUSTERING ALGORITHM

Require: D = {x1,x2,x3,……..,xn} be the set of data points

step 1: Initialization: The initial cluster centers „k‟ randomly selected from given data set D.

step 2: Distance Calculation: Calculate the distance between every data object and centroids.

step 3: Data point could be assigned to the cluster center whose distance is minimum of all the other cluster centers.

step 4: Centroid Recalculation: Recalculate the new cluster center.

step 5: The distance of each data point is again calculated with the new cluster centers.

step 6:Convergence condition: Repeat step 2 to 5 until convergence.

ENHANCED K-MEANS USING LINEAR DATA STRUCTURE LIST

In traditional K-means clustering algorithm, the distance between the data objects and the cluster centroids are calculated in each iteration which potentially affects the efficiency of clustering [Sheeba et al. 2012]. To avoid this issue, this work proposes an enhanced K-means clustering algorithm that incorporates a linear data structure list. The list stores the information about the cluster number of clustered data objects, centroids and their distances between the centroids in each iteration. The stored information is given as an input to the next iteration for the consecutive comparisons. Hence, the proposed algorithm reduces the execution time through less computations

of the distance of each data objects in clusters. This section presents a list-based data structure approach to enhance the efficiency of distance calculation in traditional K-means clustering method to produce the cluster.

Algorithm 2: ENHANCED K-MEANS ALGORITHM USING LINEAR DATA STRUCTURE LIST

INPUT : D= {d1, d2…dn} // containing n data objects.

K // Number of desired clusters.

OUTPUT : A set of k clusters Steps:

*1) Initially k data items are chosen from Dataset D, r*andomly

2) Calibrate the distance between every data object di(1 <=i<=n) and all k cluster centers cj(1<=j<=k) as Euclidean distance d(di , cj) and assign data object di to the nearest cluster.

3) For each data object di, find the nearest center cjand assign ata object di to cluster center cj

4) Detect the name of cluster center and the distance of data object di to the closest cluster. Then this information is stored in list Clu[ ] and the Dis[ ] separately.

Set Clu[i]=j, j is the name of nearest cluster. Set Dis[i]=d(di, cj), d(di, cj) is the Euclidean distance to the nearest center.

*5) R*ecalculate the cluster center;

6) Repeat

3.2.1 Distance Metrics Overview

Distance measures determine the way to calculate the similarity of two points and how it influences cluster shapes. Distance metrics also measures the similarity or regularity of data items [Li 2015]. Clustering techniques necessitate to specify the data are inter or intra-related with each other. The objective of metric calculation to a specific problem is to identify an appropriate distance function. The knowledge over metrics is critical in many learning tasks. Moreover, the metrics are applied in a wide range of applications as the problem with learning evolves a definite notion of distance or similarity. A metric function or distance function is a function that defines a distance between elements or objects of a set. A set with a metric is known as metric space [Jyoti et al. 2014]. This distance metric plays a very important role in clustering techniques. There are numerous distance methods that are available for clustering.

3.2.1.1 Euclidean Distance Metric

This is probably the very commonly chosen type of distance metric. Simply it is the geometric distance in the multidimensional space.The Euclidean distance metric calculates the sum of squared difference of co-ordinates between two points ‘a’ and ‘b’ with ‘k’ dimensions **[Malik et al. 2014]**. Equation denotes the formula for Euclidean distance between a point $x(x1, x2, \dots, xk)$ and a point $y(y1, y2, \dots, yk)$.

$CENT[i](1 \leq i \leq k) - (1)$

3.2.1.2 Manhattan Distance Metric

Unlike Euclidean distance metric, Manhattan calculates the difference between two points ‘a’ and ‘b’ by traversing through vertical and horizontal lines in the grid-based system [**Sinwar et al.2014**].Equation(1) denotes the distance calculation of Manhattan Distance metric.

Where ‘n’ is the number of variables, and ‘xi’ and ‘yi’ are the values of the ith variable, at points ‘x’ and ‘y’ respectively.

3.2.1.3 Chebychev Distance Metric

Chebychev Distance metric computes the absolute magnitude of the differences of coordinates of two pair of points ‘a’ and ‘b’. Chebychev distance is also called as the maximum value distance analysis. Equation (2) denotes the distance calculation ofChebychev metric.

$$l/m \sum_{i\ =1}^{m} [min\, d\ \ [minj\ d2xs, vj)]\ - (2)$$

where ‘xi’ and **‘**yj’ are the values of the ith variable at points ‘x’ and ‘y’, respectively.The Chebychev distance may be appropriate if the difference between points isreflected more by differences in individual dimensions rather than all the dimensionsconsidered together.

3.2.1.4 Minkowski Distance Metric

The Minkowski distance metric on Euclidean space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance. Equation(3) denotes the distance calculation of Minkowski Distance metric.
Where r is a parameter.
When r =1 Minkowski formula tend to compute Manhattan distance.
When r =2 Minkowski formula tend to compute Euclidean distance.
When r =∞ Minkowski formula tend to compute Supremum.

4. RESULTS AND DISCUSSIONS

In this experiment section, we have evaluated our enhanced algorithm on Iris dataset, Medical Diabetes Dataset and Soya bean Plant Dataset from the UCI repository of machine learning databases. We compared our results with original K-means algorithm it terms of time taken to build the model. Brief summary of the dataset used in the algorithm is given below.
For improved cluster analysis, Fisher‟sIiris, Soya bean and Medical Diabetes data sets are chosen and used with the existing training Dataset. The number of cluster k sets 6. Clustered

results for the enhanced k-means algorithm with the linear data structures are compared in this paper are listed in Table 2.

Table 1: CHARACTERISTICS OF DATAETS

Dataset	**No. Of Instances**	**No.Of Attributes**
Fisher’s Iris	150	4
Soya bean Plant	768	8
Medical diabetes Dataset	47	35

Table 2: COMPARISON OF VARIOUS DISTANCE METRICS COMPARISON WITH DIFFERENT CLUSTERING

Datasets	Distance metrics	Technique used		No. of outliers		Accuracy	
		Standard K-means	Enhanced K-means with data structure	Standard K-means	Enhanced K-means with data structure	Standard K-means	Enhanced K-means with data structure
Fisher‟s Iris	**Euclidean Distance Metric**	0.096	0.080	3	8	86.3	87.5
Soya bean Plant	**Manhattan Distance Metric**	0.081	0.069	6	14	78.9	82.5
Medical diabetes	**Chebychev Distance Metric**	0.097	0.081	4	9	91.3	96.54
Diggle	**MinkowskiDistance Metric**	0.092	0.070	5	13	79.3	90.50

Algorithm 1: Results of Distance based outlier removal algorithm in K-MEANS clustering

Maximum distance	0.4256
Minimum distance	1.7625
Threshold Value	1.09405
Accuracy before outlier removal	0.6719
Silhouette before outlier	0.4064
Accuracy after outlier removal	0.6860
Silhouette after outlier	0.4110

5. CONCLUSION

This article proposes a Framework for K-means clustering algorithm. The enhanced Framework preserve all important features of the traditional k-means and at the same time eliminates the possibility of formation of empty clusters and enhance the efficiency and accuracy of K-means clustering algorithm. A detailed comparison of this enhanced algorithm with the traditional k-means has been reported. Experimental results demonstrate that the enhanced clustering design is able to solve the empty cluster problem without any significant performance degradation.

6. REFERENCES

[1] Wei Li, “Modified k-means clustering algorithm”, IEEE computer society Congress on Image and Signal Processing, 2008, pp. 618-621.

[2] Jiawei Han and MichelineKamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, second Edition, 2006.

[3] Hatamlou, “In search of optimal centroids on data clustering using a binary search algorithm”, Pattern Recognit. Lett., vol. 33, (2012), pp. 1756-1760.

[4] F. Cao, J. Liang and G. Jiang, “An initialization method for the K-Means algorithm using neighborhood model”, Comput. Math. Appl., vol. 58, (2009), pp. 474-483.

[5] R. C. De Amorim and P. Komisarczuk, “On initializations for the minkowski weighted K-means”, proceedings of the 11th International Conference on Advances in Intelligent Data Analysis XI, (2012) October 25-27, 2012, Helsinki, Finland, pp. 45-55.

[6] Changqing Zhou, Dan Frankowski, Pamela Ludford, ShashiShekhar,LorenTerveen, “Discovering personally meaningful places: An interactive clustering approach”, Volume 25, Issue 3 ACM Transactions on Information Systems (TOIS), July 2007.

[7] Ran Vijay Singh and M.P.S Bhatia, “Data Clustering with Modified k-means Algorithm”, IEEE International Conference on Recent Trends in Information Technology, ICRTIT 2011, 2011, pp 717-721.

[8] AhamedShafeeq B M and Hareesha K S “Dynamic Clustering of Data with Modified k-Means Algorithm” International Conference on Information and Computer Networks, ICICN 2012, pp 221-225.

[9] Komarasamy G and Amitabh Wahi, “An Optimized kmeans Clustering Technique Using Bat Algorithm”, European Journal of Scientific Research, ISSN 1450-216X Vol.84 No.2, August 2012, pp.263 – 273.

[10] D T Pham, S SDimov, and C D Nguyen “Selection of k in k-means clustering”, Mechanical Engineering Science, 2004, pp. 103-119.

[11] R. Xu and D. Wunsch, II, “Survey of clustering algorithms”, IEEE Trans. Neural Networks., vol. 16, no. 3, 2005, pp. 645– 678.

[12] Baolin Yi, HaiquanQiao, Fan Yang, Chenwei Xu, “An Improved Initialization Center Algorithm for Kmeans Clustering,” IEEE 2010.

[13] J.JamesManoharan and Dr.S.Hari Ganesh, “Improved Kmeans Clustering Algorithm using Linear Data Structure List to Enhance the Efficiency”, International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.20 (2015).

[14] J.JamesManoharan and Dr.S.Hari Ganesh, “Initialization of Optimized K-means Centroids using Divide-andConquerMethod”,ARPN Journal,ISSN 1819-6608,vol 11,No.2,January 2016,pp -1086-1091.